View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

ScienceDirect

International Journal of Approximate Reasoning
43 (2006) 179–201

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Value versus damage of information release: A data privacy perspective ☆

Da-Wei Wang [a,b], Churn-Jung Liau [a,b,*], Tsan-sheng Hsu [a,b], Jeremy K.-P. Chen [c]

[a] *Institute of Information Science, Academia Sinica, Taipei 115, Taiwan*
[b] *Taiwan Information Security Center (TWISC), Taipei 115, Taiwan*
[c] *1 University Station C0803, University of Texas, Austin TX 78712, USA*

Received 19 September 2005; received in revised form 6 April 2006; accepted 6 April 2006
Available online 2 May 2006

Abstract

We assume that a database of personal information comprises records of individuals that contain confidential or sensitive fields. Queries about the distribution of a sensitive field within a selected population in the database can be submitted to the data center. However, the answers to the queries may leak confidential information about some individuals, even though no identification information is provided. Inspired by decision theory, we present two quantitative models for privacy protection in such a database query or linkage environment. One models the value of information from the viewpoint of the querier, while the other models the damage caused by and compensation for privacy leakage.

In both models, we define the information state by a class of probability distributions on a set of possible confidential values. These states can be modified and refined by the user's knowledge acquisition behavior. In the first model, the value of information is defined as the expected gain of the querier, and privacy is protected by imposing costs on the answers to the queries to balance any potential gain. In the second model, the safety of information is guaranteed by ensuring that anyone misusing private information must pay more compensation than the value of the possible gain.
© 2006 Elsevier Inc. All rights reserved.

☆ Some preliminary results of this paper have appeared in [10,26].

* Corresponding author. Address: Institute of Information Science, Academia Sinica, Taipei 115, Taiwan. Fax: +886 2 27824814.

E-mail addresses: wdw@iis.sinica.edu.tw (D.-W. Wang), liaucj@iis.sinica.edu.tw (C.-J. Liau), tshsu@iis.sinica.edu.tw (T.-s. Hsu), jchen@ece.utexas.edu (J.K.-P. Chen).

0888-613X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2006.04.002

Keywords: Data table; Decision logic; Information value; Privacy; Quantitative models

1. Introduction

Although privacy protection has recently received a great deal of attention in the data analysis community [1,5,13,12,2,9], there are still many technical problems that need to be addressed. The most basic problem is how to prevent unauthorized users gaining access to confidential information. In many cases, even if the confidential information is only known imprecisely, there is still a risk of privacy leakage. Thus, it is very important to have the capability of risk assessment in a model for privacy protection.

Some people or groups may benefit from privacy leakage; however, others may be harmed. The value of confidential information could be an incentive to invade a person's privacy and information brokers may therefore try to collect and sell such information for their own gain. On the other hand, it is usually difficult to estimate the damage caused by privacy leakage. However, to discourage the invasion of privacy, the damage to the victim must be appropriately compensated by the party disseminating the information. Therefore, the evaluation of the potential gain or loss caused by privacy leakage is a crucial problem in privacy protection.

In this paper, we address the problem in terms of the value of information and the damage caused by privacy leakage. We focus on the following database query environment. Private information about individuals is stored in a data center, and there are confidential or sensitive fields, as well as identification fields, in each record. Queries about the distribution of a sensitive field within a selected population in the database can be submitted to the data center, but the answers to queries may leak confidential information about individuals, even though no identification is provided.

We study two quantitative models for privacy protection in such a database query environment. One models the value of information from the viewpoint of the querier, while the other deals with the damage caused by and compensation for privacy leakage. For the former, we model the value of information as the expected gain from knowledge of the information. For the latter, the safety of data is guaranteed by ensuring that anyone disseminating private information without authorization must pay more compensation than the benefit he can derive from such behavior. In both models, we need to represent the knowledge states of a user receiving certain kinds of information. The user can change or refine his information state by obtaining certain answers to a query. Thus, we also need a formalism to represent the data to be protected and a language to describe the kinds of queries that are allowed. The data table and decision logic proposed in [19] are employed as the data representation formalism and the query language respectively.

In the remainder of the paper, we review the data table formalism and the decision logic in our application context in Section 2. The basic components of our models—the information state and knowledge acquisition behavior—are defined in Section 3. In Sections 4 and 5, respectively, we present a model for the value of information and a model for the damage caused by and compensation for privacy leakage. In Section 6, we discuss related works. Finally, in Section 7, we summarize the results of our work.

2. Data representation and query language

To state the privacy protection problem, we must first formalize data representation. The most popular form of data representation is a data table [19]. The data in many application domains, for example, medical records, financial transactions, employee data, etc., can be represented as data tables. A data table can be seen as a simplification of a relational database, since the latter usually consists of a number of data tables. A formal definition of a data table is given in [19].

Definition 1. A data table[1] is a pair $T = (U, A)$ such that

- U is a non-empty finite set of individuals, called the population or the universe,
- A is a non-empty finite set of primitive attributes, and
- every primitive attribute $a \in A$ is a total function $a : U \to V_a$, where V_a is the set of values of a, called the domain of a.

The attributes of a data table can be divided into three sets. The first contains the *key attributes*, which can be used to identify to whom a data record belongs; therefore, these attributes are always masked off in response to a query. It does not matter whether the set contains a single key or multiple keys, since they must be removed or encrypted before the data is released. As the key attributes uniquely determine an individual, we can assume that they are associated with elements in the universe U and omit them hereafter. Second, we have a set of *easy-to-know attributes*, the values of which can be discovered easily by the public. For example, in [22], it is pointed out that some attributes like birth-date, gender, ethnicity, etc., are included in some public databases, such as census data or voter registration lists. The last set of attributes is the *confidential type*, the values of which are the goals we have to protect. Sometimes, there is an asymmetry between the values of a confidential attribute. For example, if the attribute is a HIV test result, the revelation of a "+" value may cause a serious invasion of privacy, whereas it does not matter if it is known that an individual has a "−" value. For simplicity, we assume there is exactly one confidential attribute in a data table. Thus, a data table is usually written as $T = (U, A \cup \{c\})$, where A is the set of easy-to-know attributes and c is the confidential attribute.

Let $V_c = \{s_0, s_1, \ldots, s_{t-1}\}$ be the set of possible values for the confidential attribute c. It is assumed that the a priori information of a user is the probability distribution of the population on V_c. In other words, we assume that the user knows the value $\frac{|\{u \in U | c(u) = s_i\}|}{|U|}$ for all $0 \leqslant i \leqslant t-1$; therefore, he can improve his knowledge by investigating some sampled individuals of the population, or by querying the data center that stores the data table. By such investigation, the user can discover the exact value of the confidential attribute of the chosen individuals. However, a great deal of effort is necessary to conduct the investigation. On the other hand, a query may ask for the probability distribution of sensitive fields in a specific subset of the population. Once the query is correctly answered, the user not only knows the probability distribution of the specific sub-population, but also that of its complement on V_c. Thus, we need a language to specify a subset of individuals. To

[1] Also called a knowledge representation system, information system, or attribute-value system.

achieve this, we suggest the use of decision logic (DL) proposed in [19], which was originally designed to represent rules induced from a data table by data mining techniques. However, it is also perfectly suitable for querying a data table, since each formula of the logic is satisfied by some individuals in the data table.

The atomic formula of a data table $T = (U, A \cup \{c\})$ is of the form (a, v), where $a \in A$ is an easy-to-know attribute and $v \in V_a$ is a possible value of the attribute a. The well-formed formulas (wff) of the logic are then formed by the Boolean connectives negation ($\neg$), conjunction ($\wedge$), disjunction ($\vee$), and implication ($\rightarrow$):

- Each atomic formula is a wff.
- If φ is a wff, so is $\neg\varphi$.
- If φ and ψ are wffs, so are $\varphi \wedge \psi$, $\varphi \vee \psi$, and $\varphi \rightarrow \psi$.

The satisfaction relation $\models_T$ between U and the wffs is defined recursively by the following clauses:

(1) $u \models_T (a, v)$ iff $a(u) = v$,
(2) $u \models_T \neg\varphi$ iff $u \not\models_T \varphi$,
(3) $u \models_T \varphi \wedge \psi$ iff $u \models_T \varphi$ and $u \models_T \psi$,
(4) $u \models_T \varphi \vee \psi$ iff $u \models_T \varphi$ or $u \models_T \psi$,
(5) $u \models_T \varphi \rightarrow \psi$ iff $u \not\models_T \varphi$ or $u \models_T \psi$.

Intuitively, it is clear any individual satisfying (a, v) has v as the value of his attribute a.

Based on the semantics of decision logic, we define the truth set of a wff φ with respect to the data table T, denoted by $\|\varphi\|_T$, as $\{u \in U | u \models_T \varphi\}$. Thus, each wff φ specifies a subset of individuals $\|\varphi\|_T$ in the data table. When a user submits a query φ to the data center, it means that he wants to know the distribution of the sub-population $\|\varphi\|_T$ on V_c. If the query is correctly answered, by the axioms of probability, the user would also simultaneously know the distribution of the sub-population $U - \|\varphi\|_T$. In other words, a correctly answered query would partition the population into two sub-populations and the distributions of their confidential attribute values would be known. In this way, the user could subsequently query the data center to refine his knowledge regarding the distribution of the confidential attribute values in each sub-population. To model the evolution of the user's information states after different queries, we need a formal representation of those states. Section 3 is devoted to the definitions of such a representation.

Example 1. Let us use the data table in Fig. 1, which contains data about 100 individuals, to illustrate the definitions. The key attribute of the table is "ID"; the easy-to-know attributes are "Age" and "Sex"; and the confidential attribute is "Disease". The values of "Age" indicate "less than 30" (1), "between 30 and 60" (2), and "greater than 60" (3), whereas the values of "Disease" indicate"normal" (0) and "ill" (1). We group individuals with the same attribute values together. For example, the first row represents all males younger than 30 who do not have a disease. We add the last column to denote the count of individuals with a specific combination of values. In addition, (Sex, M) and $(\text{Sex}, F) \wedge (\text{Age}, 2)$ are two examples of decision logic wffs; thus we have $[\|(\text{Sex}, M)\|_T = \{u_1 - u_{20}, u_{31} - u_{50}, u_{71} - u_{85}\}$ and $\|(\text{Sex}, F) \wedge (\text{Age}, 2)\|_T = \{u_{51} - u_{70}\}$.

ID	Age	Sex	Disease	Count
$u_1 - u_{15}$	1	M	0	15
$u_{16} - u_{20}$	1	M	1	5
$u_{21} - u_{28}$	1	F	0	8
$u_{29} - u_{30}$	1	F	1	2
$u_{31} - u_{46}$	2	M	0	16
$u_{47} - u_{50}$	2	M	1	4
$u_{51} - u_{68}$	2	F	0	18
$u_{69} - u_{70}$	2	F	1	2
$u_{71} - u_{80}$	3	M	0	10
$u_{81} - u_{85}$	3	M	1	5
$u_{86} - u_{90}$	3	F	0	5
$u_{91} - u_{100}$	3	F	1	10

Fig. 1. An example of a data table.

3. The information states

Hereafter, we denote a data table as $T = (U, A \cup \{c\})$. Let $V_c = \{s_0, s_1, \ldots, s_{t-1}\}$ be the set of possible values for the confidential attribute, and let $U = \{u_1, u_2, \ldots, u_n\}$ be the set of individuals. A *logical partition* of U is a subset of DL wffs $\Pi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ such that $\|\varphi_1\|_T \cup \cdots \cup \|\varphi_m\|_T = U$ and $\|\varphi_i\|_T \cap \|\varphi_j\|_T = \emptyset$ if $i \neq j$. Each $\|\varphi_i\|_T$ is called an equivalence class of Π. A piece of information (or knowledge) known to the user is given by a logical partition of U, a set of probability distributions indexed by the wffs of the partition, and the number of investigated individuals. In the following, we use $|\varphi|$ to denote the cardinality of $\|\varphi\|_T$.

Definition 2. An information state (or a knowledge state) $\mathscr{I}$ for the set of possible confidential attribute values V_c and the set of individuals U is a triple $(\Pi, (\mu_i)_{0 \leqslant i \leqslant t-1}, (\kappa_i)_{0 \leqslant i \leqslant t-1})$, where Π is a logical partition of U, and for all $0 \leqslant i \leqslant t-1$, $\mu_i : \Pi \to [0,1]$ and $\kappa_i : \Pi \to \aleph_0$ ($\aleph_0$ denotes the set of natural numbers) are functions satisfying the following constraints for any $\varphi \in \Pi$,

(i) $\sum_{i=0}^{t-1} \mu_i(\varphi) = 1$,
(ii) $|\varphi| \cdot \mu_i(\varphi)$ is a natural number, and
(iii) $\kappa_i(\varphi) \leqslant |\varphi| \cdot \mu_i(\varphi)$.

For ease of description, we use vector notations to denote μ_i's and κ_i's. Thus, $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_{t-1})$ and $\boldsymbol{\kappa} = (\kappa_0, \ldots, \kappa_{t-1})$ denote vector mappings that can be applied to elements of Π. The result is a vector containing of the results of applying its component functions to the element. The dimension of each vector is self-evident from the context; therefore it is not explicitly specified. Based on the vector notation, information state defined above can be denoted by $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$. Let $\mathscr{I}$ be such an information state; then $(\Pi, \boldsymbol{\mu})$ is called a *partial knowledge state* compatible with $\mathscr{I}$. Note that a partial knowledge state may be compatible with various information states.

Within an information state, the user partitions the population into a number of sub-populations, and knows the probability distribution of the confidential attribute values of each sub-population. Intuitively, $\mu_i(\varphi)$ is the proportion of individuals in sub-population $\|\varphi\|_T$ who have the confidential attribute value s_i, whereas $\kappa_i(\varphi)$ is the number of investigated individuals in sub-population $\|\varphi\|_T$ who have confidential attribute value s_i. Since each DL wff φ is composed of atomic formulas with easy-to-know attributes only, it can be assumed that it would take little effort for the user to verify whether a given individual satisfies φ. Furthermore, it can also be assumed that the cardinality of the truth set of each φ is known to the public. However, note that it may sometimes be very difficult for the user to locate an individual who satisfies a specific φ among the whole population U.

A user can change his information state by his subsequent investigation of some individuals in a specific sub-population and by his queries to, and the answers obtained from, the data center. This is a process of knowledge refinement and can be modeled by the knowledge acquisition behavior as follows.

A logical partition Π_2 is a refinement of another logical partition Π_1, denoted by $\Pi_2 \sqsubseteq \Pi_1$, if for all $\varphi_2 \in \Pi_2$ there exists $\varphi_1 \in \Pi_1$ such that $\|\varphi_2\|_T \subseteq \|\varphi_1\|_T$. Clearly, if $\Pi_2 \sqsubseteq \Pi_1$, then each $\|\varphi_1\|_T$ such that $\varphi_1 \in \Pi_1$ can be written as a union of the truth sets of some wffs in Π_2.

Definition 3. Let $\mathscr{I}_1 = (\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$ and $\mathscr{I}_2 = (\Pi_2, \boldsymbol{\mu_2}, \boldsymbol{\kappa_2})$ be two information states. $\mathscr{I}_2$ is a refinement of $\mathscr{I}_1$, also denoted by $\mathscr{I}_2 \sqsubseteq \mathscr{I}_1$, if both of the following conditions are satisfied:

(1) $\Pi_2 \sqsubseteq \Pi_1$.
(2) For each $\varphi \in \Pi_1$, if $\|\varphi\|_T = \cup_{1 \leqslant i \leqslant l} \|\varphi_i\|_T$ for some set $\{\varphi_1, \ldots, \varphi_l\} \subseteq \Pi_2$, then

$$|\varphi| \cdot \boldsymbol{\mu_1}(\varphi) = \sum_{i=1}^{l} |\varphi_i| \cdot \boldsymbol{\mu_2}(\varphi_i),$$

and

$$\boldsymbol{\kappa_1}(\varphi) \leqslant \sum_{i=1}^{l} \boldsymbol{\kappa_2}(\varphi_i).$$

Note that the arithmetic operations (addition and multiplication) and comparison between vectors (and scalars) are defined as usual. For example, the addition of two vectors is carried out point-wise and results in a vector of the same dimension.

Example 2. In Example 1, we have $n = 100$ and $t = 2$. Let us now consider a logical partition $\Pi = \{\varphi_1, \varphi_2, \varphi_3\}$ such that $\varphi_i = (\text{Age}, i)$ for $i = 1, 2, 3$. Then, $\mathscr{I}_1 = (\Pi, (\mu_0, \mu_1), (\kappa_0, \kappa_1))$ is an information state for the data table in Fig. 1, where μ_0, μ_1, κ_0, and κ_1 are specified as follows:

	φ_1	φ_2	φ_3
μ_0	$\frac{23}{30}$	$\frac{17}{20}$	$\frac{1}{2}$
μ_1	$\frac{7}{30}$	$\frac{3}{20}$	$\frac{1}{2}$
κ_0	2	5	1
κ_1	0	1	1

Let us further define $\Pi' = \{\psi_0, \psi_1, \varphi_2, \varphi_3\}$ such that $\psi_0 = \varphi_1 \wedge (\text{Sex}, M)$ and $\psi_1 = \varphi_1 \wedge \neg(\text{Sex}, M)$. Obviously, Π' is a logical partition. If $\mu_0', \mu_1', \kappa_0'$, and κ_1' are specified as follows:

	ψ_0	ψ_1	φ_2	φ_3
μ_0'	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{17}{20}$	$\frac{1}{2}$
μ_1'	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{3}{20}$	$\frac{1}{2}$
κ_0'	2	0	10	1
κ_1'	0	0	2	1

then $\mathscr{I}_2 = (\Pi', (\mu_0', \mu_1'), (\kappa_0', \kappa_1'))$ is a refinement of $\mathscr{I}_1$.

In our framework, there are two kinds of knowledge acquisition action that can refine the user's information states. The first is a query action, which is represented by a wff φ in DL. The intended answer to the query is the distribution of confidential values within the selected population $\|\varphi\|_T$ in the database. The second type is an investigative action, which is specified by a wff φ and a positive integer number k, and means that the user has investigated k individuals from the set $\|\varphi\|_T$. For uniformity of representation, each knowledge acquisition action is written as $\alpha = (\varphi, k)$ for some DL wff φ and $k \geqslant 0$. When $k > 0$, it is an investigative action, but it is a query action if $k = 0$.

Definition 4. (1) A knowledge acquisition action $(\varphi, 0)$ is applicable under the information state $\mathscr{I}_1 = (\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$ and results in a state $\mathscr{I}_2 = (\Pi_2, \boldsymbol{\mu_2}, \boldsymbol{\kappa_2})$ if

(a) there exists $\varphi' \in \Pi_1$ such that $\|\varphi\|_T \subseteq \|\varphi'\|_T$,
(b) $\Pi_2 = \Pi_1 - \{\varphi'\} \cup \{\varphi, \varphi' \wedge \neg\varphi\}$,
(c) $\mathscr{I}_2$ is a refinement of $\mathscr{I}_1$,
(d) $\boldsymbol{\kappa_2}(\psi) = \boldsymbol{\kappa_1}(\psi)$ for any $\psi \in \Pi_1 - \{\varphi'\}$, and
(e) $\boldsymbol{\kappa_2}(\varphi) + \boldsymbol{\kappa_2}(\varphi' \wedge \neg\varphi) = \boldsymbol{\kappa_1}(\varphi')$.

(2) A knowledge acquisition action (φ, k), where $k > 0$, is applicable under the information state $\mathscr{I}_1 = (\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$, and $\mathscr{I}_2 = (\Pi_2, \boldsymbol{\mu_2}, \boldsymbol{\kappa_2})$ is a resultant state of the application if

(a) $\varphi \in \Pi_1$ and $k \leqslant |\varphi| - \sum_{i=0}^{t-1} \kappa_{1i}(\varphi)$,
(b) $\Pi_1 = \Pi_2$,
(c) $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$,
(d) $\boldsymbol{\kappa_2}(\psi) = \boldsymbol{\kappa_1}(\psi)$ for any $\psi \neq \varphi$, and
(e) $\sum_{i=0}^{t-1} \kappa_{2i}(\varphi) = \sum_{i=0}^{t-1} \kappa_{1i}(\varphi) + k$.

Since the goal of the user is to refine his knowledge by the queries, a rational user would pose his queries so that his knowledge would be improved by the answers to the queries. Thus, if the user's information state is $(\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$, he would pose a query about a subset of an equivalence class in Π_1, which is the requirement of Condition 1a in Definition 4. After the query has been answered, the corresponding equivalence class is partitioned into two parts—one satisfies φ and the other does not—so we have Condition 1b in Definition 4. Condition 1c in Definition 4 further requires that the answer is correct; thus, the resultant information state is a refinement of the original one. Furthermore, since the query action does not result in any new individuals being investigated, the $\boldsymbol{\kappa_2}$ function agrees with $\boldsymbol{\kappa_1}$ for the part of the population not split by the query. Meanwhile, for the split part, the total number of investigated individuals does not change, which is reflected in Conditions 1d and 1e of the definition.

In the case of an investigative action, we assume the user will only investigate the individuals in a sub-population represented by a wff in Π_1. The assumption is inessential, since, if the investigated individuals are from different sub-populations, the corresponding investigative action can be decomposed into a sequence of actions that satisfy the applicability condition. As it is assumed that the user knows the total number of individuals in $\|\varphi\|_T$ and the number of those investigated by him so far is equal to $\sum_{i=0}^{t-1} \kappa_i(\varphi)$, the number of individuals that he can investigate is at most the number of all un-investigated individuals. This is required by the applicability condition of Definition 4(2a). Conditions 2b–2d of the definition are obvious, since the values specified in the three conditions are not affected by the investigation. However, the investigation can affect the total number of investigated individuals in $\|\varphi\|_T$, which is reflected in Condition 2e.

Example 3. Continuing with Example 2, let us first apply the knowledge acquisition action $(\psi_0, 0)$ in the information state $\mathscr{I}_1$. Then, $\mathscr{I}_3 = (\Pi', (\mu_0', \mu_1'), (\kappa_0'', \kappa_1''))$ is a resultant state, where κ_0'', and κ_1'' are specified as follows:

	ψ_0	ψ_1	φ_2	φ_3
κ_0''	2	0	5	1
κ_1''	0	0	1	1

If the knowledge acquisition action $(\varphi_2, 6)$ is further applied in state $\mathscr{I}_3$, then $\mathscr{I}_2$ will be a possible next state. As the example shows, the resultant state of a query action is com-

pletely determined by the data table if the query is correctly answered. However, the resultant state of an investigative action is determined by the actual investigation results. Thus, there may be more than one possible resultant state for an investigative action.

4. The value of information

To quantitatively determine the value of information, we need a user model. Let us consider the case where the user is an agent trying to use some confidential information to aid his decision-making in a game he plays with some individuals in the population U. The agent can decide the rate he wants to charge an individual for playing the game (i.e., the admission fee). Furthermore, the rate is decided on an individual basis so that each player may be charged a different rate. if an individual agrees to play the game and pays the fee, he would have a chance to gain some reward, which would be a loss to the agent. The reward would be determined by the individual's confidential attribute value. Let r_i denote the reward of an individual with the confidential attribute value s_i for $0 \leqslant i \leqslant t-1$. Then, $\boldsymbol{\rho} = (r_0, r_1, \ldots, r_{t-1}) \in \Re^t$ represents the loss vector of the agent.

Let $\mathscr{I}_0 = (\{\top\}, \boldsymbol{\mu_0}, \boldsymbol{\kappa_0})$ be the initial information state of the user, where $\top$ denotes any tautology in DL and $\boldsymbol{\kappa_0}(\varphi) = (0, \ldots, 0)$; and let $\boldsymbol{\rho}$ be a given loss vector. The agent first decides the *base rate* of the game according to the expected loss which is derived from his initial information state, i.e., $R_0 = \boldsymbol{\rho} \cdot \boldsymbol{\mu_0}(\top)$. Thus, in the initial state, the expected payoff for the agent is zero. However, once he acquires extra information and reaches a new information state, he can utilize that information to make a profit.

We further assume that each individual will participate in the game if he is charged the base rate, but, he can refuse if the agent charges him more than the base rate. The higher the rate, the greater the likelihood the individual will refuse to play the game. If the information state is $\mathscr{I} = (\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$, where $\Pi = \{\varphi_1, \ldots, \varphi_m\}$, a reasonable decision by the agent for the rate of an individual u satisfying φ would be as follows.

(1) If u has been investigated and is known to have confidential attribute value s_i, then the most profitable decision of the agent would be to charge the individual $\max(R_0, r_i)$ so that his payoff would be $\max(R_0 - r_i, 0)$.
(2) If the individual has not been investigated, the agent knows that the probability of the confidential attribute value of u being s_i is

$$p_i(\varphi) = \frac{|\varphi| \cdot \mu_i(\varphi) - \kappa_i(\varphi)}{|\varphi| - \sum_{i=0}^{t-1} \kappa_i(\varphi)}. \tag{1}$$

In this case, the most reasonable decision by the agent would be to charge the individual $\max\left(R_0, \sum_{i=0}^{t-1} p_i(\varphi) \cdot r_i\right)$ so that his expected payoff would be $\max\left(R_0 - \sum_{i=0}^{t-1} p_i(\varphi) \cdot r_i, 0\right)$.

Thus, on average, the agent can expect the following payoff, B_φ, by playing the game with an individual satisfying φ:

$$B_\varphi = \max\left(R_0 - \sum_{i=0}^{t-1} (p_i(\varphi) \cdot r_i), 0\right) \cdot \frac{|\varphi| - \sum_{i=0}^{t-1} \kappa_i(\varphi)}{|\varphi|} + \sum_{i=0}^{t-1} \max(R_0 - r_i, 0) \cdot \frac{\kappa_i(\varphi)}{|\varphi|}. \tag{2}$$

By using the knowledge about individuals' confidential attributes, the agent can raise the rates of those who may incur a greater loss to him. Clearly, the value of the information depends on how much he can benefit by using it. The agent's expected gain from each individual is computed by

$$B_{\mathscr{I}} = \sum_{\varphi \in \Pi} B_{\varphi} \cdot \frac{|\varphi|}{|U|},$$

if he decides the rates according to the two principles above.

Example 4. The above scenario usually occurs between an insurance company and its clients. The base rate is applied to a typical client if the company does not have further information about his health condition. However, the company would raise the rates of high risk clients, so the health information of such clients would be valuable to the insurance company. To prevent the leakage of confidential information, the data center may increase the price for answering a query so that the value of information to the company is counter-balanced and there would be no incentive to continue the investigation.

The notion of the value of information has been studied extensively in decision theory [6,16]. In our model, as investigative actions are not allowed, all information states are of the form $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa_0})$; hence, $\kappa_{0i}(\varphi) = 0$ and $p_i(\varphi) = \mu_i(\varphi)$ for all $0 \leqslant i \leqslant t-1$, and $\varphi \in \Pi$. Consequently, $B_{\mathscr{I}}$ can be simplified to

$$\sum_{\varphi \in \Pi} \max(R_0 - \boldsymbol{\mu}(\varphi) \cdot \boldsymbol{\rho}, 0) \cdot \frac{|\varphi|}{|U|},$$

which is the value of partial information defined in [16] if our user model is appropriately formulated as a decision problem of the agent. While, in our case, partial information is obtained by querying the data center, another approach obtains partial information by sampling, as suggested in [16]. Though sampling is similar to investigation, the information obtained from the two kinds of action is quite different. With sampling actions, even though the chosen individuals may be thoroughly investigated, only statistical information can be derived. This is the information used to predict the status of the whole population. However, with investigative actions, the user can indeed obtain the personal information about each investigated individual, but he cannot make any statistical inference about the whole population from such information.

On the other hand, if no query actions are possible, the information states are always of the form $(\{\top\}, \boldsymbol{\mu_0}, \boldsymbol{\kappa})$. Once all individuals have been fully investigated (though this is hardly possible in any practical sense) the information state becomes a perfect state $\mathscr{I} = (\{\top\}, \boldsymbol{\mu_0}, \boldsymbol{\kappa})$, where $\kappa_i(\top) = \mu_{0i}(\top) \cdot |U|$, so $p_i(\top) = 0$ for all $0 \leqslant i \leqslant t-1$. Consequently, $B_{\mathscr{I}}$ can be simplified to

$$\sum_{i=0}^{t-1} \max(R_0 - r_i, 0) \cdot \mu_{0i}(\top),$$

which is precisely the value of perfect information defined in [16]. Thus, we have modeled the value of hybrid information in the above-defined framework.

4.1. Privacy protection by pricing mechanism

According to the above user model, the user can improve his payoff from 0 to $B_{\mathscr{I}}$ when his information state evolves from the initial state to $\mathscr{I}$. If the information is free-of-charge, the user would gladly receive it and the privacy of the individuals might be invaded. Thus, one approach to improving privacy protection is to charge for answers to queries so that the user cannot make a profit by obtaining confidential information. This can be achieved by employing a pricing mechanism in the data center. However, since the answer to a query may have different effects under different information states, the pricing mechanism must be adaptive to the query history of the user. In general, it is very difficult to design an adaptive pricing mechanism, since users may have to pay different prices for the same queries under different situations. Therefore, instead of charging each query separately, we consider a more restricted setting. Assume that each user is allowed to ask a batch of queries only once, after which he could try to perform further investigative actions, but the data center would not answer any more queries. Thus, the pricing mechanism decides the cost of each batch of queries so that the user cannot benefit from the answers.

Let $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$ be the information state of the user after a sequence of queries and follow-up investigative actions, where $\Pi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$. Since the data center has no control over how the user will use the information he receives, it can only guarantee that the cost of obtaining the information is high enough to prevent him making a profit from it, no matter what further investigative actions he undertakes. Thus, based on the partial knowledge state $\mathscr{P} = (\Pi, \boldsymbol{\mu})$ only, the data center must estimate the maximum payoff to the agent under different information states compatible with $\mathscr{P}$. Let $\mathbf{k} = (k_1, \ldots, k_m)$ be an m-tuple of non-negative integers and define

$$F_{\mathbf{k}} = \left\{ \boldsymbol{\kappa} \middle| \sum_{i=0}^{t-1} \kappa_i(\varphi_j) = k_j, \ \forall 1 \leqslant j \leqslant m \right\}$$

as the set of $\boldsymbol{\kappa}$ functions that denote the possible investigation results when a specific number of individuals has been investigated. Let $\mathscr{IS}(\mathscr{P}, \mathbf{k})$ denote the set of information states $(\mathscr{P}, \boldsymbol{\kappa})$ such that $\boldsymbol{\kappa} \in F_{\mathbf{k}}$. In other words, $\mathscr{IS}(\mathscr{P}, \mathbf{k})$ is the set of information states compatible with $\mathscr{P}$ and $\mathbf{k}$. Then, the maximal value of information obtained by the agent under $\mathscr{P}$ and $\mathbf{k}$ is defined as

$$B(\mathscr{P}, \mathbf{k}) = \max\{B_{\mathscr{I}} | \mathscr{I} \in \mathscr{IS}(\mathscr{P}, \mathbf{k})\}.$$

We further assume that a cost function $\gamma_{\text{inv}} : \Phi \times \mathscr{Z}^+ \to \Re^+$ is available to both the user and the data center, where Φ is the set of DL wffs and $\mathscr{Z}^+$ and $\Re^+$ are respectively the sets of positive integers and real numbers. The intended meaning of $\gamma_{\text{inv}}(\varphi, k)$ is the cost of the investigation of k individuals that satisfy φ. It can be assumed that γ_{inv} is a super-linear function in its second argument. Thus, when the user poses a batch of queries Q, the data center knows what the resultant partial knowledge state $\mathscr{P}$ will be once the answer is released. Therefore, the price of Q must be decided before releasing the information. The price, $price(Q)$, of the answers to the batch of queries should be decided such that

$$|U| \cdot B(\mathscr{P}, \mathbf{k}) - \sum_{i=1}^{m} \gamma_{\text{inv}}(\varphi_i, k_i) \leqslant price(Q) \tag{3}$$

holds for any $\mathbf{k}$. The lowest solution of $price(Q)$ for (3) is

$$\max_{\mathbf{k}} \left(|U| \cdot \max\{B_{\mathscr{I}} | \mathscr{I} \in \mathscr{IS}(\mathscr{P}, \mathbf{k})\} - \sum_{i=1}^{m} \gamma_{\text{inv}}(\varphi_i, k_i) \right), \tag{4}$$

where the domain of $\mathbf{k}$ is finite, since $0 \leqslant k_i \leqslant |\varphi_i|$. The maximization (4) is a typical combinatorial optimization problem. However, the practical estimation of its value may be computationally difficult due to the large number of possible tuples $\mathbf{k}$ and information states in each $\mathscr{IS}(\mathscr{P}, \mathbf{k})$. Non-classical techniques, such as genetic algorithms, might be helpful in solving the maximization.

4.2. Usefulness of information

In our pricing mechanism, the data center assumes that the user can play the above-mentioned game with all individuals in U and charge them according to the total gain he expects to achieve. However, this may be an over-estimation, since the user cannot play the game with all individuals when the population is large. To circumvent the problem, we may assume that the user must spend some resources to play the game with each individual. Let $\gamma_{\text{ply}} : \Phi \times \mathscr{Z}^+ \to \Re^+$ be another cost function such that $\gamma_{\text{ply}}(\varphi, l)$ denotes the cost of the user playing the game with l individuals that satisfy φ. Given an m-tuple of non-negative integers $\mathbf{l} = (l_1, \ldots, l_m)$ and an information state $\mathscr{I}$, define

$$B^{\mathbf{l}}_{\mathscr{I}} = \sum_{i=1}^{m} B_{\varphi_i} \cdot l_i.$$

The price in (4) can be replaced by

$$\max_{\mathbf{k},\mathbf{l}} \left(\max\{B^{\mathbf{l}}_{\mathscr{I}} | \mathscr{I} \in \mathscr{IS}(\mathscr{P}, \mathbf{k})\} - \sum_{i=1}^{m} \gamma_{\text{inv}}(\varphi_i, k_i) - \sum_{i=1}^{m} \gamma_{\text{ply}}(\varphi_i, l_i) \right), \tag{5}$$

where the domains of both $\mathbf{k}$ and $\mathbf{l}$ are restricted to $0 \leqslant k_i, l_i \leqslant |\varphi_i|$.

Intuitively, each l_i and k_j represent the *usefulness* of information. Given two equivalent classes in a logical partition, it may be easier to find potential members in one equivalence class than in the other, depending on the conditions each class satisfies. It may also be easier, and thus cost-effective, to investigate members of one equivalence class rather than the other. Although these two factors may be closely related, they are not necessarily the same.

Example 5. Using the insurance company model mentioned in Example 4, assume that the world's population is represented by all the adults in a country. One equivalence class may be defined as all the people living in the same county, while another class may be defined as all people whose weight is between 60 and 65 kg. It is easy for the first group of people to be investigated and added to a list of potential clients, but it is relatively difficult to investigate the second group.

Thus, the data center can decide the price of the answers to a batch of queries, Q, by a two-level maximization procedure based on (4) or (5). The outer level maximization would depend on the form of the cost functions γ_{inv} and/or γ_{ply}, so it is unlikely that an analytic solution could be found for it. However, the inner maximization can be decomposed into

m independent maximizations for each B_{φ_i}. More specifically, given φ and $0 \leqslant k \leqslant |\varphi|$, we have to find $\boldsymbol{\kappa}(\varphi)$ that maximizes B_φ among all $\boldsymbol{\kappa}$ satisfying $\sum_{i=0}^{t-1}\kappa_i(\varphi) = k$ and $\kappa_i(\varphi) \leqslant |\varphi| \cdot \mu_i(\varphi)$ for all $0 \leqslant i \leqslant t-1$. This, in turn, is equivalent to the following constraint optimization problem in the integer domain:

$$\begin{aligned} &\text{Maximize} \quad \max\left(R_0 - \sum_{i=0}^{t-1} \frac{n_i - x_i}{N-k} \cdot r_i, 0\right) \cdot \frac{N-k}{N} + \sum_{i=0}^{t-1} \max(R_0 - r_i, 0) \cdot \frac{x_i}{N} \\ &\text{s.t.} \qquad x_0 + x_1 + \cdots + x_{t-1} = k, \quad 0 \leqslant x_i \leqslant n_i \ (0 \leqslant i \leqslant t-1), \end{aligned} \tag{6}$$

where N and n_i correspond to $|\varphi|$ and $|\varphi| \cdot \mu_i(\varphi)$ respectively. The solution of Eq. (6) can be given by the following proposition for $k \leqslant N$. Without loss of generality, we assume $r_0 \geqslant r_1 \geqslant \cdots \geqslant r_{t-1}$ for the loss vector in the proposition.

Proposition 1. *Assume* $N = \sum_{i=0}^{t-1} n_i$.

(1) *If* $k = N$, *then the solution of Eq.* (6) *is* $x_i = n_i$ *for* $0 \leqslant i \leqslant t-1$ *and its maximum value is*

$$\sum_{i=0}^{t-1} \max(R_0 - r_i, 0) \cdot \frac{n_i}{N}.$$

(2) *If* $k < N$ *and* l *is the smallest natural number such that* $\sum_{i=0}^{l} n_i > k$, *then the solution of Eq.* (6) *is*

$$x_i = \begin{cases} n_i & \text{if } i < l, \\ k - \sum_{i=0}^{l-1} n_i & \text{if } i = l, \\ 0 & \text{if } i > l, \end{cases}$$

and its maximum value is

$$\max\left(R_0 - \sum_{i=l+1}^{t-1} \frac{n_i}{N-k} \cdot r_i + \frac{\sum_{i=0}^{l} n_i - k}{N-k} \cdot r_l, 0\right) \cdot \frac{N-k}{N} + \sum_{i=0}^{l-1} \max(R_0 - r_i, 0) \cdot \frac{n_i}{N} + \max(R_0 - r_l, 0) \cdot \frac{k - \sum_{i=0}^{l-1} n_i}{N}.$$

High risk individuals are those who may cause a greater loss to the agent. For low risk individuals, the investigation cannot improve the payoff for the agent; however, for high risk ones, the investigation can indeed reduce the agent's potential loss, because he can raise their admission fees appropriately. The more that high risk individuals are investigated, the more the agent can reduce his potential loss; therefore, the maximum payoff occurs when individuals with the highest risk are investigated first. This intuition is verified by the preceding proposition.

5. The damage caused by and compensation for privacy leakage

To model the damage caused by and compensation for privacy leakage, we consider another game played between an agent, called the accuser, and the individuals in U.

The accuser tries to disseminate the confidential information of the individuals. Assume $(d_0, d_1, \ldots, d_{t-1}) \in \Re^t$ and $(c_0, c_1, \ldots, c_{t-1}) \in \Re^t$ are respectively the damage and compensation vectors of the game. The rules of the game are as follows: If an individual is accused of s_i and he actually has the confidential attribute value s_i, then the damage to him is d_i, which is also the reward of the accuser. However, if he is accused of s_i and his confidential attribute value is not s_i, then he receives compensation c_i from the accuser. Thus, if $\mathscr{I} = (\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$ is an information state, the agent who accuses an un-investigated individual satisfying $\varphi \in \Pi$ of s_i would risk losing

$$L_i(\varphi) = (1 - p_i(\varphi))c_i - p_i(\varphi)d_i, \tag{7}$$

where $p_i(\varphi)$ is as defined by (1). The task of privacy protection, therefore, is to make the dissemination of individuals' confidential information unprofitable for the accuser. This is achieved by raising the accuser's expected loss to a threshold, which should be high enough to deter any rational agent risking such a loss. However, for ease of presentation, we assume the threshold is zero. Thus, an information state $\mathscr{I} = (\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$ is said to be *safe* if $L_i(\varphi) \geqslant 0$ for all $\varphi \in \Pi$ and $0 \leqslant i \leqslant t - 1$.

Example 6. Assume a person has faulty genes that increase the risk of developing some rare disease. This person's job application may be rejected, since a prospective employer may feel he is more likely to become ill and his work performance would probably suffer. By using the above information, the employer could discriminate against, and therefore harm, the applicant.

5.1. The basic model

An agent's query can only be answered if his information state will be safe after receiving the answer. It can be seen that an information state $\mathscr{I} = (\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$ is safe if $p_i(\varphi) \leqslant \frac{c_i}{c_i + d_i}$ for any $\varphi \in \Pi$ and $0 \leqslant i \leqslant t - 1$. However, since $p_i(\varphi)$ not only depends on $\mu_i(\varphi)$, but also on how many individuals have been investigated by the user, the data center cannot decide if answering a query will lead to a safe state. To guarantee the safety of an information state, the data center can use a worst-case analysis. Assume that for each wff φ, the user can investigate at most K_φ individuals in $|\varphi|$ at an affordable cost. Then, given a partial knowledge state $\mathscr{P} = (\Pi, \boldsymbol{\mu})$, which results from the answer to a query, the data center can guarantee the safety of the answer, no matter which (affordable) investigation is made by the user, as long as the following condition holds for all $\varphi \in \Pi$ and $0 \leqslant i \leqslant t - 1$:

$$\frac{|\varphi| \cdot \mu_i(\varphi)}{|\varphi| - K_\varphi} \leqslant \frac{c_i}{c_i + d_i}, \tag{8}$$

since, by (1),

$$p_i(\varphi) \leqslant \frac{|\varphi| \cdot \mu_i(\varphi)}{|\varphi| - K_\varphi}.$$

Condition (8) can be rewritten as

$$\mu_i(\varphi) \leqslant \frac{c_i}{c_i + d_i} \cdot \left(1 - \frac{K_\varphi}{|\varphi|}\right). \tag{9}$$

Let us consider some cases where Eq. (9) can be satisfied.

(1) If no investigative actions are possible, i.e., $K_\varphi = 0$, then (9) is satisfied if $\mu_i(\varphi) \leqslant \frac{c_i}{c_i+d_i}$. In such a case, if $d_i = 0$ or $c_i \gg d_i$, then the information state is still safe, even though $\mu_i(\varphi)$ is approximately equal to 1. This means that the revelation of an individual being s_i will not harm that individual, or the compensation will be sufficiently large to cover the damage caused to him. Hence, it does not matter if the user can almost certainly discover that a class of individuals has s_i value. On the other hand, if $c_i \approx d_i > 0$, then the information state is only safe when $\mu_i(\varphi)$ is less than 0.5. In other words, if the compensation cannot cover the damage sufficiently, then the user should not be allowed to know a confidential value with a certainty above 0.5.

(2) If investigation is allowed for at most K_φ individuals, then, to maintain safety, the upper bound of $\mu_i(\varphi)$ is discounted by the ratio $1 - \frac{K_\varphi}{|\varphi|}$. The discount effect is alleviated when $|\varphi| \gg K_\varphi$; thus, the larger the size of $\|\varphi\|$, the greater the possibility of achieving the safety requirement. This corresponds to the k-anonymity requirement for privacy protection in [24].

Based on the safety criterion, the data center can decide whether the user's query should be answered or refused. However, note that (9) is only a sufficient condition for the safety of data release, so we may not have to test it for every i and φ. In particular, if $d_i = 0$, then $L_i(\varphi) \geqslant 0$ holds no matter how the investigative actions are carried out, so we only have to test (9) for those i such that $d_i > 0$.

An alternative approach is to use a pricing mechanism to discourage the user. To formulate this approach, we need another cost function $\gamma_{\text{acc}} : \Phi \times \mathscr{Z}^+ \to \Re^+$, where $\gamma_{\text{acc}}(\varphi, k)$ denotes the cost to the user for accusing k individuals that satisfy φ. The minimum loss the user may incur under the partial knowledge state $\mathscr{P} = (\Pi, \boldsymbol{\mu})$ should then be

$$L^*(\varphi) = \min_{\mathbf{k},\mathbf{l}} \left[\sum_{i=0}^{t-1} L_i(\varphi, \mathbf{k}) \cdot l_i + \gamma_{\text{inv}}\left(\varphi, \sum_{i=0}^{t-1} k_i\right) + \gamma_{\text{acc}}\left(\varphi, \sum_{i=0}^{t-1} l_i\right) \right],$$

where

$$L_i(\varphi, \mathbf{k}) = c_i - (c_i + d_i) \cdot \frac{|\varphi| \cdot \mu_i(\varphi) - k_i}{|\varphi| - \sum_{i=0}^{t-1} k_i}$$

is the result of substituting (1) into (7) when $\kappa_i(\varphi) = k_i$ for $0 \leqslant i \leqslant t-1$; and the minimization has taken over all $k_i \leqslant |\varphi| \cdot \mu_i(\varphi)$ for $0 \leqslant i \leqslant t-1$ and l_i such that $\sum_{i=0}^{t-1} l_i \leqslant |\varphi|$. Then, if the answer to a batch of queries results in a partial knowledge state $(\Pi, \boldsymbol{\mu})$, its price should be determined by $\sum_{\varphi \in \Pi} price(\varphi)$, where the price of each φ is in accordance with the following equation:

$$price(\varphi) = \begin{cases} -L^*(\varphi), & \text{if } L^*(\varphi) < 0 \\ 0, & \text{otherwise.} \end{cases}$$

5.2. The extended model

In the basic model, we assume the damage vector is associated with each specific value of the confidential attribute. This means that if an individual is known to have the

attribute value s_i, then the harm caused to him will be d_i. However, sometimes it is harmful if an individual's attribute value is known to be in some specific subset of V_c, even if the subset is not a singleton.

Example 7. Assume a disease is classified in stages according to a range 0–5, where 0 means no disease, 1–3 can be cured, and 4–5 are terminal. Then, the knowledge that a person has been diagnosed as stage 4 or stage 5 could be harmful to that person if it became widely known.

The basic model can be extended to cover such a case. Since it is reasonable that the compensation should be in proportion to the amount of damage caused, we can simplify the model by assuming that there is a function $\alpha : \Re \to \Re$ that maps each damage value to a corresponding level of compensation. For example, it may be the case that $\alpha(x) = r \cdot x$ for some positive number r. Thus, in the extended model, we can concentrate on estimating the damage. In the model, we assume there is a damage function $\delta : (2^{\{0,\ldots,t-1\}} - \{\emptyset\}) \to \Re$. For any $S \subseteq \{0,\ldots,t-1\}$, $\delta(S)$ means the damage to an individual when it is known that his confidential attribute value belongs to $\{s_i | i \in S\}$. Using the same game rules as the basic model, the expected loss to the agent accusing an individual in $\varphi \in \Pi$ of $\{s_i | i \in S\}$ would be

$$L_S(\varphi) = \left(1 - \sum_{i \in S} p_i(\varphi)\right)\alpha(\delta(S)) - \left(\sum_{i \in S} p_i(\varphi)\right)\delta(S),$$

where $p_i(\varphi)$ is defined by (1). Then, the safety criterion for an information state $\mathscr{I} = (\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$ can be extended to

$$L_S(\varphi) \geqslant 0$$

for all $\varphi \in \Pi$ and $S \subseteq \{0,\ldots,t-1\}$. This is equivalent to

$$\sum_{i \in S} p_i(\varphi) \leqslant \frac{\alpha(\delta(S))}{\alpha(\delta(S)) + \delta(S)}. \tag{10}$$

By using the same worst-case analysis as in the basic model, the following must be satisfied for all $\varphi \in \Pi$ and $S \subseteq \{0,\ldots,t-1\}$,

$$\sum_{i \in S} \frac{|\varphi| \cdot \mu_i(\varphi)}{|\varphi| - K_\varphi} \leqslant \frac{\alpha(\delta(S))}{\alpha(\delta(S)) + \delta(S)},$$

or alternatively in the following form:

$$\sum_{i \in S} \mu_i(\varphi) \leqslant \left(\frac{\alpha(\delta(S))}{\alpha(\delta(S)) + \delta(S)}\right) \cdot \left(1 - \frac{K_\varphi}{|\varphi|}\right). \tag{11}$$

So far, the model has not addressed the estimation of the damage function δ. In fact, as the damage vector in the basic model should be determined by some external mechanism, such as the legal system or social conventions, we can assume that $\delta(\{i\}) = d_i$ is given for each $0 \leqslant i \leqslant t-1$. However, for other values of the damage function, it should be possible to impose some reasonable constraints so that the damage caused by a subset S can be (partially) estimated by the damage caused by its elements. These conditions include

(1) $\delta(\{0,\ldots,t-1\}) = 0$,
(2) $\delta(S_1) \leqslant \delta(S_2)$ if $S_2 \subseteq S_1$,

(3) $\delta(S) = 0$ if $|S| > 1$, and
(4) $\delta(S) = \min_{i \in S} \delta(\{i\})$.

Note that we do not require that all these conditions hold simultaneously. In general, Condition 1 should hold for all damage functions; however, Conditions 2–4 are optional.

Condition 1 means that if there is no privacy leakage, there is no damage. Since it is known that all possible values of the confidential attribute are in V_c, the index set $\{0, \ldots, t-1\}$ corresponds to the state of no privacy leakage. Condition 2 means the more specific the information known, the greater the damage that is likely to be caused. Condition 3 corresponds to the basic model in which only the damage value of the singleton is considered. Condition 4 is based on the principle of least commitment, which implies that if an individual is accused of a set of possible faults disjunctively, it can only be assumed that he has the least harmful fault. Therefore, the damage caused to him by such an accusation would be equivalent to the minimal damage caused by accusing him of a fault in the set. Note that Conditions 3 and 4 are not compatible if there are at least two indices, i and j, such that $\delta(\{i\}) > 0$ and $\delta(\{j\}) > 0$. However, both Conditions 1 and 2 are implied by Condition 3, and Condition 4 implies Condition 2. Furthermore, Condition 4 also implies Condition 1, provided that i exists such that $\delta(\{i\}) = 0$.

An alternative way to estimate the damage value of a subset is by the information-theoretic approach. In other words, if the a priori probability function of the possible values of the confidential attribute is given by $\boldsymbol{\mu}(\top)$, then we can compute the *a posteriori* probability for any $S \subseteq \{0, 1, \ldots, t-1\}$ as

$$Pr(s_i|S) = \begin{cases} \frac{\mu_i(\top)}{\sum_{j \in S} \mu_j(\top)}, & \text{if } i \in S; \\ 0, & \text{otherwise.} \end{cases}$$

Then, a possible constraint on the damage function is

$$\delta(S) = \sum_{i \in S} m(\mu_i(\top), Pr(s_i \mid S)) \cdot \delta(\{i\}), \tag{12}$$

where $m : [0,1] \times [0,1] \to [0,1]$ is called an information distance function. The function estimates how much the user's information about some specific s_i would be increased by knowing that its index is in S. Typically, the information distance function can be defined as the relative difference between the entropy values of the two probabilities, i.e.,

$$m(p, q) = \frac{\log p - \log q}{\log p}.$$

If the information distance function is defined in this way, then (12) can be reduced to

$$\delta(S) = \log p_S \cdot \sum_{i \in S} \frac{\delta(\{i\})}{\log p_i}, \tag{13}$$

where $p_i = \mu_i(\top)$ and $p_S = \sum_{i \in S} \mu_i(\top)$. Consequently, of the four conditions mentioned above, only Condition 1 is satisfied by the entropy-based damage function.

6. Related works

Although quantifying the value of information is by no means a novel problem, our quantitative models for privacy protection provide a new perspective on the matter. As

shown in Section 4, our model generalizes a standard notion in decision theory [16,6]. While decision-theoretic analysis [16] emphasizes the value of information from the decision-maker's viewpoint, our model is primarily concerned with privacy protection by the information provider. For the former, a decision-maker can decide if he purchases a piece of information according to the value of the information. For the latter, the information provider can charge the user of the information an appropriate fee.

An alternative model for measuring the value of information in the context of privacy protection is proposed in [2–4,25]. In the model, the value of information is estimated according to the expected cost incurred by the user to achieve a perfect knowledge state from the given information. More specifically, let a piece of information be a pair $(\varphi,(n_i)_{0\leqslant i\leqslant t-1})$ such that φ is a DL wff and $|\varphi|=\sum_{i=0}^{t-1}n_i$. The information $(\varphi,(n_i)_{0\leqslant i\leqslant t-1})$ means that, in the sub-population $\|\varphi\|_T$, there are exactly n_i individuals with confidential attribute value s_i for $0\leqslant i\leqslant t-1$. To know the confidential values of all individuals satisfying φ, the user can investigate them one by one. Assume the cost of investigating each individual is fixed, then the total cost is proportional to the number of individuals the user must investigate to know all individuals' confidential values. Though the maximum number of investigative actions the user must perform may be equal to $|\varphi|$, a piece of information $(\varphi,(n_i)_{0\leqslant i\leqslant t-1})$ may help him reduce the investigation cost. An extreme case would be when $n_i=|\varphi|$ and $n_j=0$ for all $1\leqslant j\neq i\leqslant t-1$. In such a case, the user would not have to carry out an investigation since he would know that all individuals have the confidential value s_i. Though the model in [2–4,25] is restricted to the case where the domain of values for the confidential attribute has exactly two elements, the result can be generalized according to the following proposition.

Proposition 2. *Given a piece of information* $(\varphi,(n_i)_{0\leqslant i\leqslant t-1})$, *the expected number of investigative actions required to know the confidential values of all individuals satisfying* φ *is*

$$\sum_{i=0}^{t-1}\frac{n_i\cdot(|\varphi|-n_i)}{|\varphi|-n_i+1}. \tag{14}$$

The value of information by Eq. (14) is based on the rationale that the more a piece of information can be used to reduce the investigation effort, the more valuable it is. Thus, the value of information can be defined as a monotonically decreasing function of (14). However, without the user model, the value of information based on (14) may not reflect the real situation. For example, (14) is obviously invariant under the permutation of n_i. In other words, if $\sigma:\{0,\ldots,t-1\}\to\{0,\ldots,t-1\}$ is a permutation of the index set, then

$$\sum_{i=0}^{t-1}\frac{n_i\cdot(|\varphi|-n_i)}{|\varphi|-n_i+1}=\sum_{i=0}^{t-1}\frac{n_{\sigma(i)}\cdot(|\varphi|-n_{\sigma(i)})}{|\varphi|-n_{\sigma(i)}+1}.$$

This means that two pieces of information, $(\varphi_1,(n_i)_{0\leqslant i\leqslant t-1})$ and $(\varphi_2,(n_{\sigma(i)})_{0\leqslant i\leqslant t-1})$, have the same value if $|\varphi_1|=|\varphi_2|$. However, in practice, knowing that most individuals have confidential value s_0 may have a different value than knowing that the same number of individuals have confidential value s_1.

Example 8. Let the confidential attribute denote a HIV test result and $V_c=\{+,-\}$ (i.e., $t=2$). Then, given two sub-populations of size 10,000 characterized by φ_1 and φ_2

respectively, two pieces of information $(\varphi_1,(9999,1))$ and $(\varphi_2,(1,9999))$ should have different values in terms of privacy protection, since the former says that most individuals in the sub-population are infected, whereas the latter says the opposite. Our model deals with the problem by imposing different loss values, r_0 and r_1, on the two test results.

Besides decision-theoretic analysis, the value of information can also be estimated by some information theoretic measures. The central notion of such measures is the entropy introduced by Shannon [21]. In machine learning literature, it is used to define the information gain of an attribute for a classification problem [17]. By reformulating Shannon's notion to fit our context, the entropy of a partial knowledge state $\mathscr{P} = (\Pi, \boldsymbol{\mu})$ is defined by

$$H(\mathscr{P}) = \sum_{\varphi\in\Pi} \frac{|\varphi|}{|U|} \cdot \sum_{i=0}^{t-1} \mu_i(\varphi) \cdot \log \frac{1}{\mu_i(\varphi)}.$$

The information gain derived from a partial knowledge state is defined as the difference between its entropy and that of the initial information state, i.e.,

$$Gain(\mathscr{P}) = H(\mathscr{P}_0) - H(\mathscr{P}),$$

where $\mathscr{P}_0 = (\{\top\}, \boldsymbol{\mu}_0)$ is the partial knowledge state compatible with the initial information state. Though information gain is a useful index for selecting the most informative features for the classification problem, the definition of entropy does not consider the asymmetric sensitivity of confidential attribute values.

In contrast to the quantitative approach of this paper, some qualitative criteria for privacy protection have been proposed in [9,11,20,24,23]. These criteria are designed to protect sensitive personal information in a released microdata set, i.e., a set of records containing information about individuals. The main objective is to prevent the re-identification of individuals; that is, to prevent the possibility of deducing which record corresponds to a particular individual, even though the explicit identifier of the individual is not contained in the released information. In contrast, our models are concerned with the release of statistical information, which is generally less specific than microdata. However, microdata release can also be handled by our framework when the queries are sufficiently specific. Let us define a complete specification formula (CSF) as a DL wff of the form $\wedge_{a\in A}(a, v_a)$, where A is the set of all easy-to-know attributes and v_a is a value in the domain of A. The answer to a batch of queries Q consisting of all CSF's is equivalent to releasing the microdata of the whole data table T. Therefore, our models are applicable in a more general context.

Irrespective of different application contexts, our models are still comparable with qualitative models. In the description of the μ-ARGUS system [11], it is emphasized that re-identification of an individual can occur if the individual has an easy-to-know attribute value that is rare in the population. In our notation, this means that when a query φ is posed, if $|\varphi|$ is small, then it is rather unsafe to answer the query. In particular, if $|\varphi| = 1$, then the answer to the query necessarily results in the re-identification of the individual satisfying φ. This intuition is formally investigated in [24,23]. In that framework, a formal requirement, called k-anonymity, is defined, and generalization and suppression techniques are employed to ensure that the requirement is satisfied. Generally speaking, k-anonymity requires that each bin must contain at least k individuals, where a bin is defined as an equivalence class of individuals who have exactly the same easy-to-know

attribute values. These safety criteria are also related to the granular computing paradigm [12,15,18,25]. The k-anonymity requirement can be easily enforced in our model, if it is restricted so that a query φ cannot be answered if the size $|\varphi|$ is smaller than some threshold. However, instead of generalizing or suppressing the data, we try to assess its potential value or the damage that would be caused by its release, and discourage misuse of the information by some pricing or penalty mechanism.

In Section 5.1, we noted that, when the size of $|\varphi|$ is large enough, the safety requirement for answering the query φ in our model can be achieved more easily. This also indirectly validates why k-anonymity is useful for privacy protection. Indeed, the diversity of confidential attribute values tends to be higher in a larger (sub)population. However, theoretically speaking, k-anonymity does not fully exclude the possibility of privacy leakage. Imagine a case where all individuals in the bin have the same confidential value. The sensitive information of the individuals would be simultaneously leaked if the data were to be released, even though the bin size criterion is satisfied. To circumvent the problem, a logical criterion is proposed in [2–4,9,25]. It is based on the possible world semantics of epistemic logic [7], so it is possible to rigorously define what a user (or an intruder) knows. The release of data is then said to be (logically) safe if it does not result in the confidential information being known by the user. The quantitative models defined in this paper can be seen as generalizations of the logical model. Let us temporarily leave aside the investigative actions and consider an information state $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa_0})$ such that $\kappa_{0i}(\varphi) = 0$ for all $\varphi \in \Pi$ and $0 \leqslant i \leqslant t-1$. If for some $\varphi \in \Pi$, we have $\mu_i(\varphi) = 1$ and $\mu_j(\varphi) = 0$ for $j \neq i$, then according to (7), the information state is unsafe if $d_i > 0$. This corresponds to the case in the logical model where the user knows that all individuals satisfying φ have the confidential value s_i. In addition, since $d_i > 0$, the information is sensitive, so the information state is also unsafe according to the logical criterion. However, even though the user cannot know any individual's sensitive information with certainty, the knowledge of the distribution of the confidential attribute values within a group of individuals may increase the risk of privacy invasion. Our quantitative models improve this aspect of the logical criterion. Moreover, the k-anonymity requirement can still be considered as complementary to the logical criterion, since the latter does not consider the size of the population, which is another aspect handled by the investigative actions in our models. Both the k-anonymity requirement and the logical criterion are implemented in the Cellsecu system [3,25].

Example 9. Let us reconsider the confidential attribute of the HIV test result, where $V_c = \{+, -\}$. Assume the damage caused by and compensation for the "+" result being known are respectively r and $5r$ for some $r > 0$, and that the user can investigate at most one individual at an affordable cost. By substituting d_i and c_i in (9) with r and $5r$ respectively, the safety condition is $\mu_0(\varphi) \leqslant \frac{5}{6} \cdot (1 - \frac{1}{|\varphi|})$. Now, for a partial knowledge state $\mathscr{P} = (\{\varphi, \neg\varphi\}, \boldsymbol{\mu})$, if $|\varphi| = |\neg\varphi| = 2$ and $\mu_0(\varphi) = \mu_0(\neg\varphi) = \frac{1}{2}$, then by the logical criterion, $\mathscr{P}$ is a safe knowledge state, even though it is unsafe according to the 3-anonymity requirement. Since $\mu_0(\varphi) = \frac{1}{2} > \frac{5}{12} = \frac{5}{6} \cdot \left(1 - \frac{1}{|\varphi|}\right)$, the unsafe state can be detected by the quantitative criterion proposed in this paper.

On the other hand, if $|\varphi| = |\neg\varphi| = 10$, but $\mu_0(\varphi) = 1$, then $\mathscr{P}$ satisfies the k-anonymity requirement up to $k \leqslant 10$, even though the logical criterion of safety is obviously violated. The violation of logical safety can still be detected by the quantitative criterion, since $\mu_0(\varphi) = 1 > \frac{3}{4} = \frac{5}{6} \cdot \left(1 - \frac{1}{|\varphi|}\right)$.

7. Conclusion

In this paper, we have presented two quantitative models for privacy protection. In both, a formal representation of the user's information states is given. In the first model, we estimate the value of information to the user by considering a specific user model. Under this model, the objective of privacy protection is to ensure that a user cannot profit from obtaining confidential information. It must be emphasized that the value of information is defined in terms of this particular user model. When other user models are considered, the value of information may be different. Some examples can be found in [14].

In the second model, we assume that the damage caused by and compensation for revealing each specific confidential value is known. An information state is safe when there is only a small probability that a user could discover a confidential value that would cause a large amount of damage if it were revealed.

A problem with the pricing mechanism arises naturally, since different users may put different values on the same information. This means that we may have to set different prices for different users of the same information. However, this is not as odd as it seems at first. In fact, different pricing structures are already applied in the software market, usually for educational and commercial uses. Of course, this also means that more experimental studies are needed before the models can be put to practical use. In [3,25], a prototypical system has been implemented to test the effectiveness of several safety criteria for privacy protection. In the future, we will extend the system to conduct experiments on the models proposed in this paper.

There are other complicated problems in privacy protection that cannot be resolved by purely technical means. For example, our schemes cannot prevent a group of users from collectively investigating confidential information by individually querying a data center. This must be considered from a legal standpoint. Upon releasing data to a user, the user must be required to sign a contract prohibiting him from revealing the data to others. The possibility of collusion by a group of users would thus be prohibited by law. In future works, we will investigate how technology and privacy laws can be fully combined to protect privacy.

Appendix A. Proof of Proposition 2

Let us define a $(\varphi,(n_i)_{0\leqslant i\leqslant t-1})$-*trace* (or simply a trace) as a string of symbols in V_c, the domain of values for the confidential attribute, where s_i occurs n_i times for $0 \leqslant i \leqslant t-1$ and set $N = |\varphi|$. A trace is a possible result when the user investigates the individuals in $\|\varphi\|_T$ one by one. If the last k symbols of a trace are all s_i for some i and the symbol before the last k symbols is s_j for some $j \neq i$, then the number of investigative actions the user has to perform is in fact $N-k$, since after the $(N-k)$th individual has been investigated, the user knows the remaining individuals have the confidential value s_i according to the information. Let us call such trace a (k,s_i)-trace. For each i and $0 < k \leqslant n_i$, there are in total

$$\sum_{j\neq i} \frac{(N-k-1)!}{\prod_{l\neq i,j} n_l! \cdot (n_j-1)! \cdot (n_i-k)!} \tag{A.1}$$

(k, s_i)-traces. Since $N = \sum_{i=0}^{t-1} n_i$, (A.1) can be rewritten as

$$\sum_{j \neq i} \frac{(N-k-1)! \cdot n_j}{\prod_{l \neq i} n_l! \cdot (n_i - k)!} = \frac{(N-k-1)! \cdot (N-n_i)}{\prod_{l \neq i} n_l! \cdot (n_i - k)!}. \tag{A.2}$$

Because the total number of possible traces is

$$\frac{N!}{\prod_{l=0}^{t-1} n_l!}, \tag{A.3}$$

the probability of a trace being a (k, s_i)-trace is the division of (A.2) by (A.3), i.e.,

$$\frac{(N-k-1)! \cdot n_i! \cdot (N-n_i)}{N! \cdot (n_i - k)!}. \tag{A.4}$$

Thus, the expected number of investigative actions the user has to perform to discover all individuals' confidential values is equal to

$$\sum_{i=0}^{t-1} \sum_{k=1}^{n_i} \frac{(N-k-1)! \cdot n_i! \cdot (N-n_i)}{N! \cdot (n_i - k)!} \cdot (N-k), \tag{A.5}$$

which can be further rewritten as

$$\sum_{i=0}^{t-1} \left(\sum_{k=1}^{n_i} \frac{(N-k)! \cdot n_i!}{N! \cdot (n_i - k)!} \right) \cdot (N - n_i). \tag{A.6}$$

According to [8] (pp. 173–174, Problem 1),

$$\sum_{k=0}^{n_i} \frac{\binom{n_i}{k}}{\binom{N}{k}} = \frac{N+1}{N-n_i+1},$$

so

$$\sum_{k=1}^{n_i} \frac{(N-k)! \cdot n_i!}{N! \cdot (n_i - k)!} = \frac{n_i}{N-n_i+1};$$

therefore, (A.6) can be simplified into

$$\sum_{i=0}^{t-1} \frac{n_i \cdot (N-n_i)}{N-n_i+1}, \tag{A.7}$$

which is exactly (14), since $|\varphi| = N$. □

References

[1] D. Agrawal, C.C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms, in: Proceedings of the 12th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 2001, pp. 247–255.

[2] Y.-C. Chiang, T.-s. Hsu, S. Kuo, D.-W. Wang, Preserving confidentially when sharing medical data, in: Proceedings of Asia Pacific Medical Informatics Conference, 2000.

[3] Y.C. Chiang, T.-s. Hsu, S. Kuo, C.J. Liau, D.W. Wang, Preserving confidentiality when sharing medical database with the Cellsecu system, International Journal of Medical Informatics 71 (2003) 17–23.
[4] Y.T. Chiang, Y.C. Chiang, T.-s. Hsu, C.J. Liau, D.W. Wang, How much privacy? – a system to safe guard personal privacy while releasing database, in: Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing, LNCS, vol. 2475, Springer-Verlag, 2002, pp. 226–233.
[5] C. Clifton, M. Kantarcıoğlu, J. Vaidya, Privacy-preserving data mining, in: W.W. Chu, T.Y. Lin (Eds.), Foundations and Advances in Data Mining, Springer-Verlag, 2005, pp. 313–344.
[6] G.D. Eppen, F.J. Gould, Quantitative Concepts for Management, Prentice-Hall, 1985.
[7] R. Fagin, J.Y. Halpern, Y. Moses, M.Y. Vardi, Reasoning about Knowledge, MIT Press, 1996.
[8] R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics: A Foundation for Computer Science, Addison-Wesley, 1994.
[9] T.-s. Hsu, C.-J. Liau, D.-W. Wang, A logical model for privacy protection, in: Proceedings of the 4th International Conference on Information Security, LNCS, vol. 2200, Springer-Verlag, 2001, pp. 110–124.
[10] T.-s. Hsu, C.J. Liau, D.W. Wang, Jeremy K.P. Chen, Quantifying privacy leakage through answering database queries, in: Proceedings of the 5th International Conference on Information Security, LNCS, vol. 2433, Springer-Verlag, 2002, pp. 162–175.
[11] A.J. Hundepool, L.C.R.J. Willenborg, μ- and τ-argus: Software for statistical disclosure control, in: Proceedings of the 3rd International Seminar on Statistical Confidentiality, 1996.
[12] S. Im, Z.W. Raś, Ensuring data security against knowledge discovery in distributed information systems, in: Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, LNCS, vol. 3642, Springer-Verlag, 2005, pp. 548–557.
[13] V. Iyengar, Transforming data to satisfy privacy constraints, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, 2002, pp. 279–288.
[14] J. Kleinberg, C.H. Papadimitriou, P. Raghavan, On the value of private information, in: Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge, 2001.
[15] T.Y. Lin, Granular computing on binary relations I: Data mining and neighborhood systems, in: A. Skoworn, L. Polkowski (Eds.), Rough Sets in Knowledge Discovery Vol. 1: Methodology and Applications, Physica-Verlag, 1998, pp. 107–121.
[16] D.V. Lindley, Making Decisions, John Wiley & Sons, 1985.
[17] T. Mitchell, Machine Learning, McGraw-Hill, 1997.
[18] A. Øhrn, L. Ohno-Machado, Using boolean reasoning to anonymize databases, Artificial Intelligence in Medicine 15 (1999) 235–254.
[19] Z. Pawlak, Rough Sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.
[20] P. Samarati, Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering 13 (6) (2001) 1010–1027.
[21] C.E. Shannon, The mathematical theory of communication, The Bell System Technical Journal 27 (3&4) (1948) 379–423, 623–656.
[22] L. Sweeney, Guaranteeing anonymity when sharing medical data, the datafly system, in: Proceedings of American Medical Informatics Association, 1997.
[23] L. Sweeney, Achieving *k*-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) (2002) 571–588.
[24] L. Sweeney, *k*-anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) (2002) 557–570.
[25] D.W. Wang, C.J. Liau, T.-s. Hsu, Medical privacy protection based on granular computing, Artificial Intelligence in Medicine 32 (2) (2004) 137–149.
[26] D.W. Wang, C.J. Liau, T.-s. Hsu, Jeremy K.P. Chen, On the damage and compensation of privacy leakage, in: Proceedings of the 18th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Kluwer Academic Publisher, 2004, pp. 311–324.